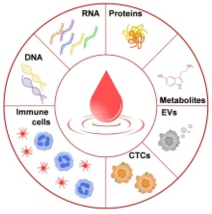
न्यूरोएंडोक्राइन ट्यूमर का कृत्रिम मेधा आधारित

9

न्यूरोएंडोक्राइन ट्यूमर का रक्त-आधारित पता लगाने में सक्षम कृत्रिम मेधा

महेश कुमार पडवाल [1,2], राहुल वी. परघाने [2,3], अविक चक्रवर्ती [2,3], भक्ति बसु [1,2]* और संदीप बसु [2,3]*

[1]आण्विक जीवविज्ञान प्रभाग, भाभा परमाणु अनुसंधान केंद्र, ट्रांबे-400085, भारत
[2]होमी भाभा राष्ट्रीय संस्थान, अणुशक्ति नगर, मुंबई-400094, भारत
[3]विकिरण औषधि केंद्र, भाभा परमाणु अनुसंधान केंद्र, टाटा स्मारक केंद्र उप भवन, परेल, मुंबई-400012, भारत



रक्तप्रवाह में ट्यूमर द्वारा जारी/संशोधित विभिन्न प्रकार के छोटे संकेतों का सचित्र चित्रण।

सारांश

मयह शोध पत्र रक्त के नमूने से न्यूरोएंडोक्राइन ट्यूमर (NET) रोग का पता लगाने से संबंधित है। यह तकनीक स्वस्थ नमूनों को न्यूरोएंडोक्राइन ट्यूमर नमूनों से अलग करने के लिए भिन्न-भिन्न अनुकूलित RNA अणुओं और कृत्रिम मेधा-आधारित गणितीय तर्कशास्त्र का उपयोग करती है। कृत्रिम मेधा -आधारित वर्गीकरण गणितीय तर्कशास्त्र ने रोग की उपस्थिति का पूर्वानुमान करने में उच्च सटीकता (>95%) प्राप्त किया है। यह उपकरण विभिन्न विशेषताओं को एकीकृत कर सकता है, जिसका न्यूरोएंडोक्राइन ट्यूमर रोगियों के व्यक्तिगत प्रबंधन के लिए विविध अनुप्रयोग हो सकता है। उपचार अनुक्रिया की निगरानी और वंशानुगत खतरे का पूर्वानुमान लगाने के लिए उपकरण की क्षमता पर चर्चा की गई है।

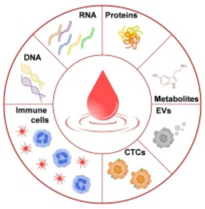
AI based Detection of Neuroendocrine Tumors

9

The Power of AI Enables Blood-based Detection of Neuroendocrine Tumors

Mahesh Kumar Padwal[1,2], Rahul V. Parghane[2,3], Avik Chakraborty[2,3], Bhakti Basu[1,2]* and Sandip Basu[2,3]*

[1]Molecular Biology Division, Bhabha Atomic Research Centre, Trombay-400085, INDIA
[2]Homi Bhabha National Institute, Anushakti Nagar, Mumbai-400094, INDIA
[3]Radiation Medicine Centre, Bhabha Atomic Research Centre, Tata Memorial Centre Annexe, Parel, Mumbai-400012, INDIA



Pictorial depiction of a variety of tiny signals released/ modulated by the tumor in the bloodstream

ABSTRACT

This communication deals with the detection of the Neuroendocrine Tumor (NET) disease from the blood sample. The technique utilizes a set of differentially expressed RNA molecules and an AI-based algorithm to distinguish healthy samples from the NET samples. The AI-based classification algorithm achieved high accuracy (> 95%) in predicting the presence of the disease. The tool can integrate various features that may have diverse applications for the personalized management of NET patients. The potential of the tool for monitoring treatment response and predicting inherited susceptibility is discussed.

KEYWORDS: *Neuroendocrine Tumors, Peripheral blood transcriptome, Machine learning, Blood biomarkers*

**Authors for Correspondence: Bhakti Basu & Sandip Basu
E-mail: bbasu@barc.gov.in & drsanb@barc.gov.in*

Introduction

Why blood biomarkers?

The incidence and prevalence of cancer are increasing by the day and so is the race to detect this dreadful disease early [1-2]. As India transitions to becoming the most populous country in the world, there are high odds of a steep disease burden. Early detection warrants an easy, non-invasive, sensitive (fewer false negatives) and specific (fewer false positives) test that uses a surrogate sample representative of a disease state. Depending on the location of the cancer, several body fluids are potential candidates as the sample source. However, being in touch with the entire body, blood has gained widespread attention and research focus for the detection of disease-specific biomarkers [3]. Moreover, blood analysis also offers insight into the genetic susceptibility to the disease [4]. This emerging tool, called *Liquid biopsy* in comparison to conventional invasive tissue biopsy, is currently the focus of active research the world over [3].

Types of biomarkers

A drop of blood contains millions of cells and billions of biomolecules in equilibrium with the health status of an individual. Blood is a carrier of various biomolecules, vesicles, etc. along with RBCs, WBCs, and platelets. In cancer patients, the cancerous tissue releases tiny signals in the form of circulating tumor DNA (ctDNA), circulating tumor RNA (ctRNA), protein markers, metabolites, extracellular vesicles, and circulating tumor cells (CTCs) into the bloodstream (Fig.1). The evidence suggests that the tumors *educate* certain immune cells and can dysregulate the systemic immune system [5]. If these tiny tumor-specific signals can be detected, quantified, distinguished from the healthy samples, and correlated with the clinical parameters of the patient, it is possible to devise a blood test to diagnose, monitor treatment response, and predict the prognosis of cancer.

Detecting disease-specific biomarkers in blood - A formidable challenge

Routine blood tests to profile levels of sugar, lipids, minerals, vitamins, hormones, parasites, etc. measure monoanalyte markers and associate the marker with the underlying disease condition, e.g., high level of blood sugar (> 120 mg/dL) is associated with diabetes. However, as cancer is a complex disease that involves local and systemic changes, monoanalyte markers often demonstrate low accuracy. The multi-analyte markers are identified in 2 ways. One of the ways is to identify the tumor-specific markers in tumor tissues by comparison with the normal tissues. Such tumor-specific markers represent the local changes in the tumor [6], which can be detected in the blood by qTR-PCR. The second way is to identify the disease-specific markers by directly profiling the blood samples from the cancer patients and comparing them with those of the healthy donors. These markers capture the local as well as the systemic changes induced by the tumor. In the case of DNA or RNA markers, the sequencing method can fish out pathogenic inherited/ sporadic germ-line mutations that impart susceptibility to early disease onset. Here, we present our recent findings on the peripheral blood RNA-sequencing-based multi-analyte tool for the diagnosis of Neuroendocrine Tumors (NETs).

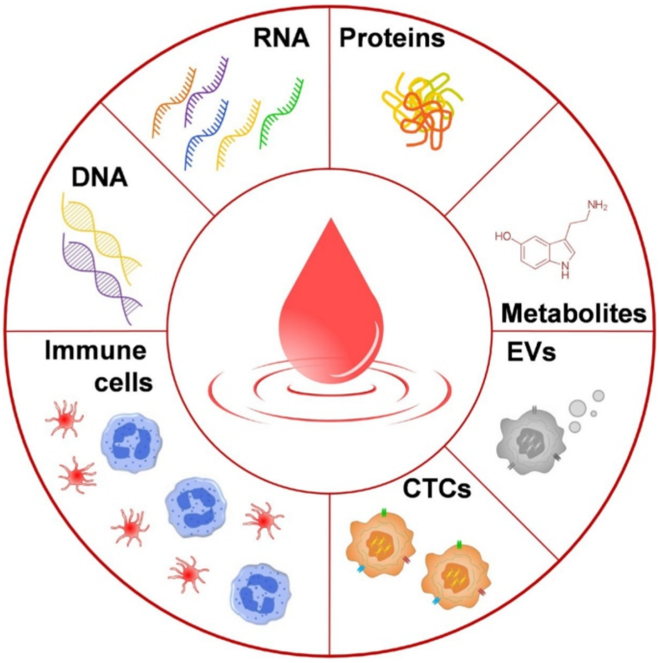


Fig.1: Pictorial depiction of a variety of tiny signals released/ modulated by the tumor in the bloodstream.

Materials and Methods

Study participants

The study participants included healthy donors (n = 51) and NET patients (n = 86) registered for ^{177}Lu-DOTATATE PRRT at Radiation Medicine Centre, India, according to eligibility criteria detailed earlier [7]. Written informed consent was obtained from all the participants and the study was approved by the Institutional Scientific Advisory Committee (SAC) and Institutional Ethics Committee (IEC). All NET samples were collected before the first PRRT cycle. The blood samples were collected in BD Vacutainer® K2 EDTA Tubes and were processed for RBC lysis by hypotonic shock method. The intact blood cells were lysed in TRI reagent (Sigma) and RNA was isolated using RNeasy Mini Kit (QIAGEN). RNA libraries were prepared with Ultra II Directional RNA-Seq Library Prep kit (NEB) and sequenced on Illumina HiSeq X instrument at the CAP-accredited commercial facility of Medgenome Lab Ltd., Bangalore, India.

Processing of RNA-Seq data

Clean RNA-Seq reads obtained using FASTP [8] were mapped to the human reference genome (GRCh38) using a splice-aware STAR aligner [9]. Gene counts were quantified using RSEM [10]. Differential expression analyses were performed on gene counts using the DESeq2 package [11]. The training set and test set samples were normalized separately. Low expression and highly variant genes were removed. The remaining 17220 genes were VST normalized and adjusted for batch effects using SVA.

Feature selection and diagnostic classifier

Relevant and complementary features were selected from VST-normalized and SVs-adjusted counts of 17,220 genes from the training sample set using the mRMRe R package [12]. Random forest algorithm was developed using the CARET package and multi-analyte gene features. Hyper-parameter mtry was optimized and the RF model was cross-validated using a 5-fold cross-validation method repeated 10 times. The performance metrics of the diagnostic classifier were assessed based on sensitivity, specificity, and accuracy.

Results and Discussion

Fig.2 shows an overall study design which includes data from 107 samples in the training set and 30 samples in the test set. The training set samples comprised NETs of the pancreas, gastrointestinal tract, and lung. The NET patients had either

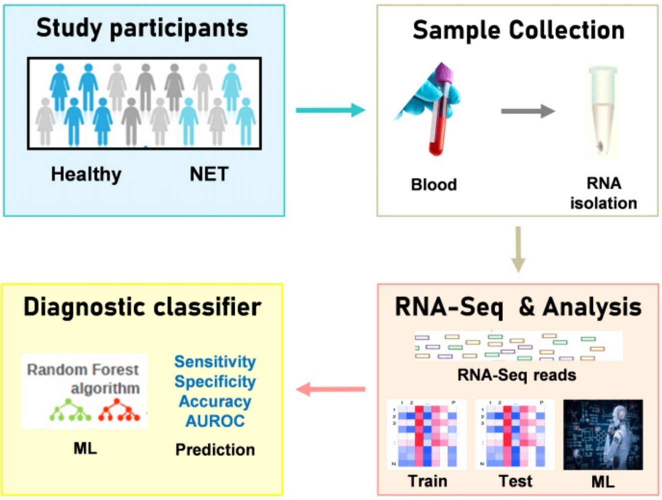


Fig.2: Study design.

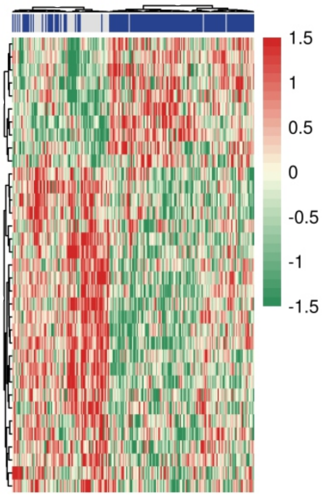


Fig.3: The hierarchical heatmap shows the segregation of samples of NET patients and healthy donors based on the expression of the most relevant gene features. The color bar represents Log2 fold change in the expression levels between the patients' samples (blue bar at the top) and the samples from the healthy donors (grey bar at the top).

localized or metastatic disease. About 40% of the NET patients had hormonal syndrome. More than 94% of RNA-Seq reads aligned to the human genome, and the reads comprised protein-coding genes, long non-coding RNAs (lncRNAs), and other small RNAs. The peripheral blood transcriptome was enriched with the genes encoded by the immune cells as well as the neuroendocrine cells. The differentially expressed genes in the blood of NET patients were used to develop a diagnostic classifier.

In the training set samples, RNA-Seq data revealed 1500 differentially expressed genes. The selection of the NET-specific gene features was guided by the expression levels of each gene and lower variability among the samples. Expression values of the selected gene features grouped the patients' samples and the healthy samples into different clusters, with some overlap (Fig.3).

The expression values of the selected NET-specific gene features were used as input features to train a diagnostic classifier using a random forest algorithm. Random forest is a decision tree algorithm that partitions the sample into either healthy or NET according to the expression levels of the NET-specific gene features (Fig.4 a). Five hundred such decision trees determine the final prediction of the sample.

In the training set, the classifier achieved 100% sensitivity and 100% specificity (Fig.4 b). In the test set, the classifier achieved 94.4% sensitivity and 100% specificity (Fig.4 b). In addition to expression-based prediction, qualitative data on RNA sequencing is useful for the assessment of pathogenic gene mutations potentially associated with the disease.

Conclusion

We have demonstrated an advanced tool for cancer detection from a blood sample. The method can be used to diagnose NET cases with high accuracy and has a potential clinical utility in monitoring the treatment response. Integration of mutation data on genes associated with NET would enable the identification of cases with inherited susceptibility to the disease or predict prognosis when combined with the relevant clinical parameters and imaging modalities. Taken together, the tool offers scope for personalized management of Neuroendocrine Tumor patients. To the best of our knowledge, this is the first report on RNA-Seq-based blood test for cancer in India.

Acknowledgment

The authors gratefully acknowledge the study participants for consenting to donate blood samples for this study. They also acknowledge Mrs. Lalita Gangurde, and Mrs. Anita Shinde, RMC for their technical support in blood collection.

References

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, Cancer Statistics 2022, CA Cancer J Clin., 2022, 72, 7-33.

[2] K. Sathishkumar, M. Chaturvedi, P. Das, S. Stephen, P. Mathur, Cancer incidence estimates for 2022 & projection for 2025:

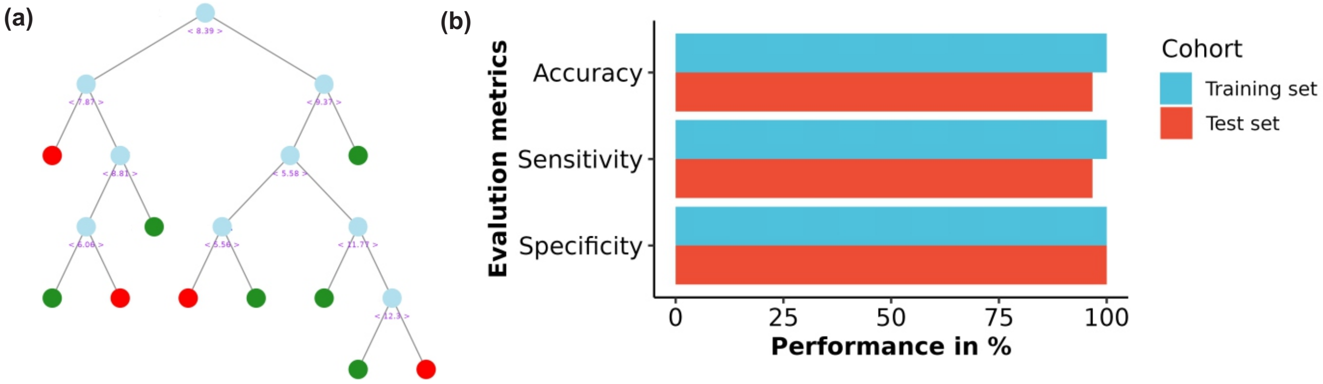


Fig.4: Machine learning to identify NETs from the blood sample. (a) A representative decision tree of the diagnostic classifier. Expression values are indicated (< value >) below each gene feature (light blue circle). Red and green circles represent the prediction of NET and healthy, respectively. (b) Accuracy metrics for detecting NETs in training and the test sets.

Result from National Cancer Registry Programme, India, Indian J Med Res., 2022, 156, 598-607.

[3] J. Marrugo-Ramírez, M. Mir, and J. Samitier, Blood-Based Cancer Biomarkers in Liquid Biopsy: A Promising Non-Invasive Alternative to Tissue Biopsy, Int J Mol Sci., 2018, 19, 2877.

[4] B. Basu and S. Basu, Correlating and combining Proteomic assessment with in vivo Molecular functional Imaging: Will this be the future roadmap for Personalized Cancer management? Cancer Biother. Radiopharm., 2016, 31, 75-84.

[5] K. J. Hiam-Galvez, B. M. Allen, and M. H. Spitzer, Systemic immunity in cancer. Nat Rev Cancer, 2021, 21, 345-359.

[6] M. K. Padwal, S. Basu, and B. Basu, Application of machine learning in predicting hepatic metastasis or primary site in gastroenteropancreatic neuroendocrine tumors. Current Oncology, 2023, 30, 9244-9261.

[7] K. Sitani, R.V. Parghane, S. Talole, and S. Basu, Long-term outcome of indigenous (177)Lu-DOTATATE PRRT in patients with Metastatic Advanced Neuroendocrine Tumours: A single institutional observation in a large tertiary care setting, Br. J. Radiol., 2021, 94, 20201041.

[8] S. Chen, Y. Zhou, Y. Chen, and J. Gu, fastp: An ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, 2018, 34, 884-890.

[9] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner, Bioinformatics, 2013, 29, 15-21.

[10] B. Li, and C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics, 2011, 12, 323.

[11] M. I. Love, W. Huber, and S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol., 2014, 15, 550.

[12] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, mRMRe: An R package for parallelized mRMR ensemble feature selection, Bioinformatics, 2013, 29, 2365-2368.